

Analyzing Digital Image by Deep Learning for Melanoma Diagnosis

Karl Thurnhofer-Hemsi, Enrique Domínguez

Department of Computer Languages and Computer Sciences.
University of Málaga.
Boulevard Louis Pasteur, 35. 29071 Málaga. Spain.
{karlkhader,enriqued}@lcc.uma.es

Abstract. Image classification is an important task in many medical applications, in order to achieve an adequate diagnostic of different lesions. Melanoma is a frequent kind of skin cancer, which most of them can be detected by visual exploration. Heterogeneity and database size are the most important difficulties to overcome in order to obtain a good classification performance. In this work, a deep learning based method for accurate classification of wound regions is proposed. Raw images are fed into a Convolutional Neural Network (CNN) producing a probability of being a melanoma or a non-melanoma. Alexnet and GoogLeNet were used due to their well-known effectiveness. Moreover, data augmentation was used to increase the number of input images. Experiments show that the compared models can achieve high performance in terms of mean accuracy with very few data and without any preprocessing.

Keywords: Image Processing, Deep Learning, Melanoma

1 Introduction

Skin cancer was the most commonly diagnosed cancer in the United States in 2016. Melanoma accounts for only 1% of all skin cancer cases, but the vast majority of skin cancer deaths. This is a type of skin cancer caused by abnormal multiplication of pigment producing cells that give color to the skin: melanocytes.

Melanoma is highly curable when detected in its earliest stages, it is more likely than other skin cancer to spread to other parts of the body [1]. Melanoma, in their initial growth phases, and other benign moles are similarities in their characteristic, which makes the diagnosis difficult between what is malignant and what is benign for experienced dermatologists [8].

The best way to detect skin cancer early is to recognize new or changing skin growths, particularly those that look different from other moles. The ABCDE rule (Fig.1) outlines warning signs of the most common type of melanoma: A is for asymmetry, B is for border irregularity, C is for color, D is for diameter greater than 6 millimeters and E is for evolution [11].

Many previous techniques in dermatological computer aided classification has lacked the generalization capability of medical practitioners due to insufficient data and require extensive preprocessing [2], lesion segmentation [6, 7] and

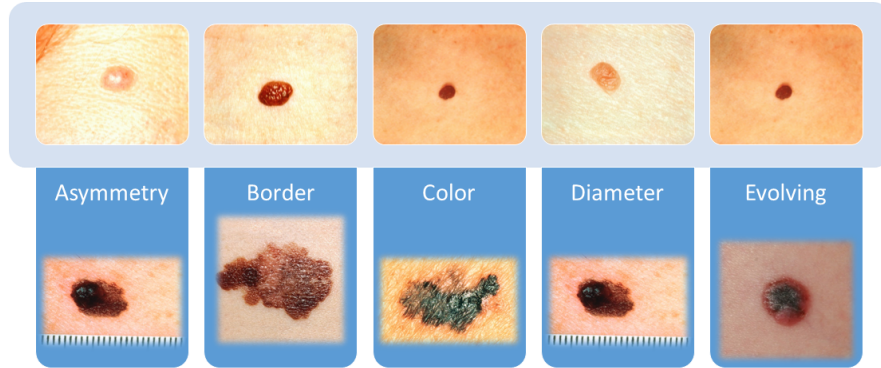


Fig. 1. Traditional clinical analysis followed by dermatologists (ABCDE rule).

extraction of specific features before classification [4, 16]. Several learning-based methods were applied form melanoma classification [13, 3, 15] using the special features of this kind of lesion. By contrast, our proposal requires no human intervention, since it is trained directly from the labelled images with a single neural network.

Convolutional Neural Networks (CNNs) is one of the most popular deep learning techniques for image analysis. CNNs were inspired by the animal visual cortex and they are one of the first truly successful deep learning architectures, which have shown outstanding performance in processing images and videos. Nowadays, with the help of GPU-accelerated computing techniques, CNNs have been successfully applied to object recognition (e.g. handwriting, face, behavior...), recommender systems or image classification. Recent works show that deep networks are being a powerful tools for medical image analysis [10], and therefore for melanoma classification [12, 17].

In this work, two different deep learning based methods have been implemented on a computer for detection of melanoma lesions, which could assist a dermatologist in early diagnosis of this cancer. Both methods were evaluated on the provided images from the DermIS and DermQuest databases (now integrated in Derm101¹). The whole dataset is composed by 533 images, where 329 are melanomas and 204 nevi. The provided experimental results are really promising, even more if we think about the few images that have been used. Deep learning often requires hundreds of thousands or millions of images for the best results.

The rest of the paper is organized as follows: Section 2 describes the deep learning techniques used in this work. Experiments and discussed results are

¹ Comprehensive digital resource for healthcare professionals including information from world-renowned experts in the field. <https://www.derm101.com/>

reflected in Section 3. Finally, conclusions and future works are summarized in Section 4.

2 Deep Learning Techniques

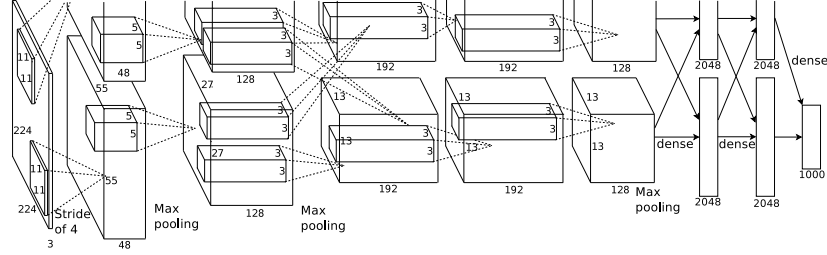
Deep learning algorithms, powered by advances in computation and very large datasets, have shown to exceed human performance in visual tasks such as playing games (Chess, Go, etc.) and object recognition. In this work, we outline a development of a CNN in order to assist and improve the performance of dermatologists in melanoma image based detection.

Machine learning technologies have already achieved significant success in many areas including classification, regression and clustering. However, many of these techniques work well only under a common assumption the the training and test data are drawn from the same feature space and the same distribution. Most statistical models need to be retrained or rebuilt with newly collected data when the distribution changes. Unfortunately, in many real world applications this process could be expensive or simply impossible. In such cases, transfer learning can truly be suitable and beneficial.

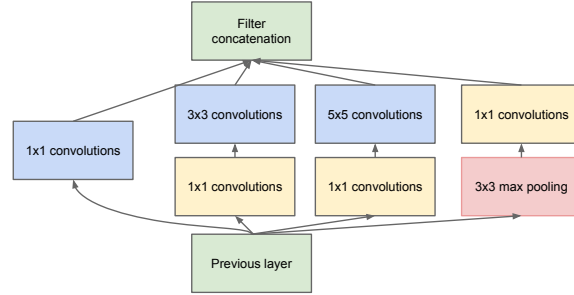
The need for transfer learning may arise when the data is easily outdated or when the distribution of data is very different among the distinct classes. In such situations, it is easier to adapt a classification model, which has been trained on a similar problem, to help learn classification models for equal nature data. There exists a large amount of works on transfer learning for reinforcement in the literature. Nevertheless, in this paper, we only focus on transfer learning for classification problems by using two widely known CNNs, such as AlexNet and GoogLeNet, that are related more closely to the addressed problem.

AlexNet [9] consists of five convolutional layers and three fully connected layers, that won the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) in 2012. A schematic draw of the net's architecture is shown in Fig. 2a. The input is an RGB image of size 256×256 , and it corresponds to one of 1000 different classes. The output is a vector of probabilities with one value for each class. Random crops of size 227×227 are generated from inside the 256×256 images to feed the first layer of AlexNet. The first two convolutional layers are followed by the max pooling layers. The third, fourth and fifth convolutional layers are connected directly. The last is followed by a maxpool, and connected to a series of two fully connected layers, to end into a softmax classifier with 1000 class labels. ReLU activation is applied after all the convolution and fully connected layers.

GoogLeNet [14] is a deep neural network architecture for computer vision codenamed Inception. It was presented at the ILSVRC 2014 classification and detection challenges, being one of the best the state-of-the-art deep networks. The philosophy of this model is the deeper the better. This consists on both increasing the depth of the network and the number of neurons at each layer. It is composed by a total of 22 layers, most of them based on Inception modules as depicted in Fig. 2b. All the convolutions use ReLU activation. The input of



(a) Alexnet (image extracted from [9])



(b) Inception of GoogLeNet (image extracted from [14])

Fig. 2. Architecture of the networks.

the network is 224×224 taking RGB color channels with mean subtraction. The output, as Alexnet, is a vector of 1000 components to determine between 1000 classes.

Both networks were trained on 1.2 million images from the ImageNet database². We used these pre-trained models and trained them on our dataset using transfer learning in order to transfer the knowledge of these models to our target problem.

A typical problem of these deep learning techniques are the requirement of large datasets to perform a good model training. Small datasets provoke deep networks do not generalize well data from the test set. Thus, these networks suffer from the problem of over-fitting. Data augmentation is a simply way to reduce over-fitting on models. The amount of training data is increased using only its information. Data augmentation techniques have commonly been applied to medical classification problems during last years with success [5]. Depending on the problem we are dealing with, different augmentation methods can be applied. The main techniques are based on data warping, which is an approach which augment the input data in data space. A commonly practice for augmenting image data is to perform geometric deformations, such as reflections, crops and translations of the image. Also, changing the color and luminosity of the image

² <http://www.image-net.org>

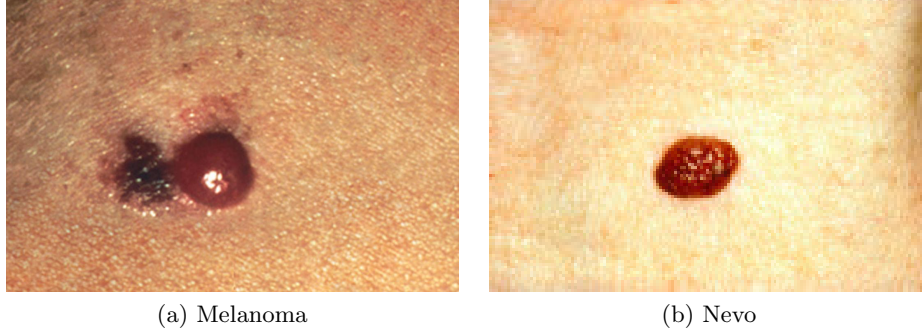


Fig. 3. Example of the two different classes of images.

is useful in some cases. All of them are performed as an affine transformation of the original image.

3 Experimental results

In this section are summarized the experiments we carried out in order to evaluate the performance of the proposed models.

A total of 533 images, where 62% belongs to the melanoma class and the rest are classified as nevi, were used to train and test the models (Fig. 3). It is a small dataset for a deep learning model, but one of our objectives is to show that a high accuracy can be achieved with few samples. For this purpose, data augmentation techniques are employed for precision improvement. These are methods for constructing iterative sampling algorithms via the introduction of unobserved data. In these experiments, we carried out reflections, crops and rotations to the data.

There are several parameters that can be tuned in a deep neural network. However, the main ones we are focusing on are:

- Batch Size (BS): indicates the amount of images processed in one iteration.
- Initial learning rate (LR): establishes with which rate is going to start the learning procedure.
- Validation frequencies (Val. Freq.): is the number of iterations between evaluations of the validation metrics.

The parameter values selected for analysis are summarized in Table 1.

From a quantitative point of view, we have chosen a well-known measure in order to compare the performance of the detection of the different studied networks: the Accuracy (ACC). It measures the proportion of true detections among the total number of tested images.

$$ACC = \frac{TP + TN}{TP + TN + FN + FP} \quad (1)$$

Table 1. Parameters for neural network tuning

Parameter	Tested values
BS	{16, 32, 64, 128}
LR	{0.0001, 0.00001, 0.000001}
Val. Freq.	{5, 10, 15, 20}

where TP , TN , FP , and FN denotes the true positives, true negatives, false positives and false negatives, respectively.

All the possible combinations of the values mentioned in Table 1 were used for training and the outcomes of the test. In Table 2 are summarized the outcomes of the proposed models, using both raw dataset and data augmentation.

With the non use of data augmentation, GoogLeNet performs slightly better than Alexnet in average. Using batch sizes 16 or 32, there exist learning rates with which the mean accuracy of Alexnet overcome GoogleNet, specially using validations frequencies 15 and 20. However, for higher batch sizes and independently from the learning rate used, GoogLeNet obtains the best results with a best accuracy of almost 88%.

The best configurations are summarized in Table 3. When we use the original dataset without applying data augmentation, smaller learning rates perform better, whereas that when the data is augmented more stability is achieved with a high learning rate for both networks. The optimal validation frequency for any model is almost the same, which would say that this parameter is not too relevant. On the other hand, the batch size values are discrepant and symmetric with respect the data augmentation technique: Alexnet generates the best accuracy with a batch size of 128 with more data and GoogLeNet needs 16, but when the original dataset is used, GoogLeNet carried on a better optimization with 128 and Alexnet with 32, reflecting their dependencies on the dataset size.

Fig. 4 shows the training loss and accuracy of the best configuration of the Alexnet network using data augmentation. The smoothness of the plots indicates that the model learned properly, reducing the fluctuations while the number of epochs increased. The loss stabilization indicates that no over-fitting has occurred and the absence of significant drops in the accuracy curve suggest that the model is robust. This panorama changes when data augmentation is employed during the training phase. Despite the fact that both methods improve substantially their results, now Alexnet is the best model for almost any configuration of the parameters, depicting an accuracy of 93,4% of detections in front of the best 92% of detections of the GoogLeNet.

Finally, average confusion matrices for the best configurations are presented in Table 4. The outcomes of the raw dataset (Tables 4a and 4c) are worse than the augmented one. The amount of false negatives is around 8-9% for both Alexnet and GoogLeNet, which is still quite high. Using Alexnet, only 52% of the dataset is detected as melanoma (in front of the 62% that are present on the dataset), and results reach 54% of true positives for GoogLeNet. The best rate of false negatives (lower is better) is achieved using data augmentation (Tables

Table 2. Average accuracy and standard deviation of the proposed models. Best results are boldfaced.

BS	LR	Val. Freq.	Non-Augmented		Augmented	
			Alexnet	GoogleNet	Alexnet	GoogleNet
16	1E-06	5	0,839 \pm 0,06	0,845 \pm 0,02	0,927 \pm 0,02	0,886 \pm 0,05
		10	0,834 \pm 0,03	0,845 \pm 0,06	0,927 \pm 0,01	0,906 \pm 0,03
		15	0,843 \pm 0,05	0,831 \pm 0,05	0,928 \pm 0,01	0,897 \pm 0,07
		20	0,835 \pm 0,05	0,837 \pm 0,04	0,931 \pm 0,01	0,896 \pm 0,02
	1E-05	5	0,835 \pm 0,04	0,859 \pm 0,04	0,930 \pm 0,01	0,897 \pm 0,06
		10	0,826 \pm 0,05	0,826 \pm 0,04	0,923 \pm 0,01	0,914 \pm 0,03
		15	0,826 \pm 0,04	0,841 \pm 0,03	0,930 \pm 0,01	0,886 \pm 0,04
		20	0,811 \pm 0,03	0,857 \pm 0,06	0,926 \pm 0,01	0,912 \pm 0,02
	1E-04	5	0,842 \pm 0,06	0,847 \pm 0,04	0,923 \pm 0,01	0,912 \pm 0,03
		10	0,831 \pm 0,04	0,837 \pm 0,06	0,927 \pm 0,01	0,909 \pm 0,02
		15	0,816 \pm 0,09	0,828 \pm 0,03	0,928 \pm 0,01	0,920 \pm 0,02
		20	0,836 \pm 0,04	0,864 \pm 0,03	0,925 \pm 0,02	0,895 \pm 0,03
32	1E-06	5	0,841 \pm 0,04	0,857 \pm 0,03	0,924 \pm 0,01	0,889 \pm 0,03
		10	0,820 \pm 0,04	0,802 \pm 0,07	0,930 \pm 0,01	0,896 \pm 0,03
		15	0,839 \pm 0,03	0,819 \pm 0,05	0,930 \pm 0,01	0,903 \pm 0,02
		20	0,833 \pm 0,03	0,833 \pm 0,04	0,929 \pm 0,01	0,888 \pm 0,03
	1E-05	5	0,833 \pm 0,03	0,839 \pm 0,05	0,925 \pm 0,01	0,901 \pm 0,03
		10	0,842 \pm 0,05	0,868 \pm 0,05	0,925 \pm 0,03	0,907 \pm 0,02
		15	0,834 \pm 0,04	0,856 \pm 0,07	0,927 \pm 0,01	0,894 \pm 0,02
		20	0,851 \pm 0,03	0,842 \pm 0,05	0,926 \pm 0,01	0,911 \pm 0,02
	1E-04	5	0,809 \pm 0,06	0,830 \pm 0,03	0,929 \pm 0,01	0,905 \pm 0,02
		10	0,841 \pm 0,06	0,837 \pm 0,05	0,926 \pm 0,01	0,914 \pm 0,01
		15	0,832 \pm 0,04	0,837 \pm 0,07	0,922 \pm 0,01	0,904 \pm 0,02
		20	0,848 \pm 0,02	0,856 \pm 0,07	0,928 \pm 0,01	0,910 \pm 0,03
64	1E-06	5	0,827 \pm 0,05	0,834 \pm 0,04	0,930 \pm 0,01	0,879 \pm 0,05
		10	0,835 \pm 0,08	0,847 \pm 0,02	0,925 \pm 0,01	0,896 \pm 0,06
		15	0,843 \pm 0,04	0,829 \pm 0,04	0,927 \pm 0,01	0,905 \pm 0,07
		20	0,826 \pm 0,05	0,841 \pm 0,06	0,929 \pm 0,00	0,904 \pm 0,04
	1E-05	5	0,830 \pm 0,04	0,841 \pm 0,04	0,923 \pm 0,01	0,917 \pm 0,03
		10	0,817 \pm 0,05	0,826 \pm 0,03	0,928 \pm 0,01	0,872 \pm 0,04
		15	0,826 \pm 0,06	0,857 \pm 0,02	0,928 \pm 0,01	0,904 \pm 0,04
		20	0,808 \pm 0,05	0,847 \pm 0,04	0,924 \pm 0,01	0,910 \pm 0,05
	1E-04	5	0,805 \pm 0,04	0,836 \pm 0,06	0,921 \pm 0,03	0,900 \pm 0,04
		10	0,832 \pm 0,04	0,868 \pm 0,05	0,925 \pm 0,01	0,908 \pm 0,05
		15	0,841 \pm 0,02	0,831 \pm 0,08	0,925 \pm 0,01	0,915 \pm 0,03
		20	0,826 \pm 0,07	0,834 \pm 0,05	0,924 \pm 0,01	0,906 \pm 0,03
128	1E-06	5	0,835 \pm 0,05	0,829 \pm 0,05	0,924 \pm 0,01	0,913 \pm 0,04
		10	0,842 \pm 0,02	0,841 \pm 0,05	0,928 \pm 0,01	0,906 \pm 0,03
		15	0,841 \pm 0,04	0,878 \pm 0,04	0,932 \pm 0,01	0,906 \pm 0,03
		20	0,831 \pm 0,05	0,822 \pm 0,07	0,919 \pm 0,02	0,906 \pm 0,03
	1E-05	5	0,841 \pm 0,07	0,854 \pm 0,05	0,925 \pm 0,01	0,899 \pm 0,04
		10	0,849 \pm 0,04	0,856 \pm 0,05	0,925 \pm 0,01	0,906 \pm 0,05
		15	0,827 \pm 0,04	0,856 \pm 0,05	0,927 \pm 0,03	0,893 \pm 0,03
		20	0,825 \pm 0,05	0,812 \pm 0,06	0,927 \pm 0,02	0,903 \pm 0,06
	1E-04	5	0,823 \pm 0,04	0,857 \pm 0,03	0,929 \pm 0,01	0,912 \pm 0,02
		10	0,831 \pm 0,04	0,854 \pm 0,04	0,923 \pm 0,01	0,917 \pm 0,01
		15	0,841 \pm 0,05	0,854 \pm 0,02	0,934 \pm 0,01	0,908 \pm 0,02
		20	0,816 \pm 0,03	0,843 \pm 0,04	0,925 \pm 0,02	0,895 \pm 0,03

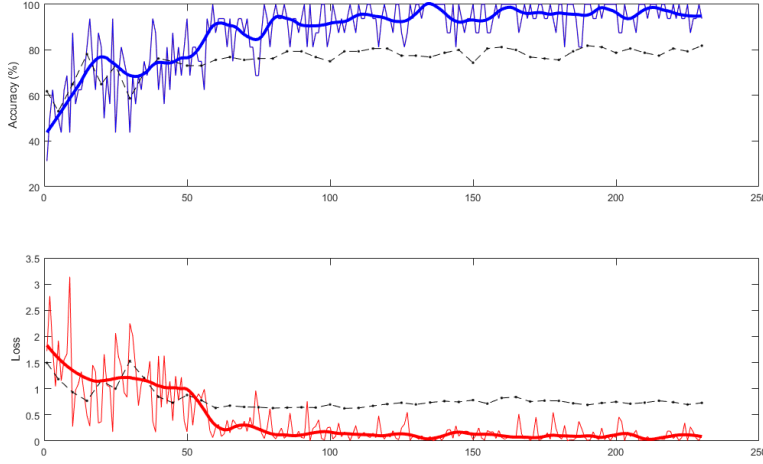


Fig. 4. Training progress of the best configuration of the augmented Alexnet.

Table 3. Best configuration of the proposed models.

Configuration	Non-Augmented		Augmented	
	Alexnet	GoogLeNet	Alexnet	GoogLeNet
Batch Size	32	128	128	16
Learning Rate	1E-05	1E-06	1E-04	1E-04
Validation Frequency	20	15	15	15

4b and 4d), with 3% of FN for Alexnet. Moreover, 58% of the melanomas are detected correctly. Although FP is not as much important as FN in medical applications, the results showed that it never surpasses 7% with raw data, which is acceptable, and this rate improves until 4% for the augmented data. The conjunction with good percentages of TP and TN demonstrate that our system would be applicable to melanoma diagnostics.

4 Conclusions

In this paper we introduced an automatic classification model for melanoma detection through convolutional neural networks. We employed transfer learning techniques with two well-known deep networks: Alexnet and GoogLeNet. The combination with data augmentation showed an improvement of the performance, producing state-of-the-art results for melanoma detection. Experiments on the DermQuest database demonstrate that 93% of accuracy can be achieved with few images, unbalanced classes and with a not very deep network, providing an effective platform for melanoma diagnostic with no human interaction and avoiding the segmentation step.

Table 4. Average matrix confusion of the proposed models for the best configuration.

		Predicted Melanoma Nevus	
Real	Melanoma	0.52	0.09
	Nevus	0.07	0.32

(a) Non-Augmented Alexnet

		Predicted Melanoma Nevus	
Real	Melanoma	0.58	0.03
	Nevus	0.04	0.35

(b) Augmented Alexnet

		Predicted Melanoma Nevus	
Real	Melanoma	0.54	0.08
	Nevus	0.07	0.31

(c) Non-Augmented GoogleNet

		Predicted Melanoma Nevus	
Real	Melanoma	0.56	0.06
	Nevus	0.04	0.34

(d) Augmented GoogleNet

Further works include the use of the other datasets, which would improve the performance and reduce the false negative detections even more. An extensive statistical analysis with more precision measures applied to a detailed optimization of the network model would allow us to achieve better results for the following cancer diagnosis and healing processes.

Acknowledgements

This work is partially supported by the Ministry of Economy and Competitiveness of Spain under grant TIN2014-53465-R, project name Video surveillance by active search of anomalous events. It is also partially supported by the Autonomous Government of Andalusia (Spain) under grant TIC-657, project name Self-organizing systems and robust estimators for video surveillance. All of them include funds from the European Regional Development Fund (ERDF). The authors thankfully the grant of the University of Málaga and acknowledge the computer resources, technical expertise and assistance provided by the SCBI (Supercomputing and Bioinformatics) center of the University of Málaga. Karl Thurnhofer-Hemsi (FPU15/06512) is funded by a PhD scholarship from the Spanish Ministry of Education, Culture and Sport under the FPU program.

References

1. American Cancer Society, I. (ed.): Cancer Facts & Figures 2016. Atlanta: American Cancer Society (2016)
2. Asha Gnana Priya, H., Anitha, J., Poonima Jacinth, J.: Identification of melanoma in dermoscopy images using image processing algorithms. In: 2018 International Conference on Control, Power, Communication and Computing Technologies, IC-CPCCT 2018. pp. 553–557 (2018)
3. Bakheet, S.: An svm framework for malignant melanoma detection based on optimized hog features. *Computation* 5(1) (2017)

4. Devassy, B., Yildirim-Yayilgan, S., Hardeberg, J.: The impact of replacing complex hand-crafted features with standard features for melanoma classification using both hand-crafted and deep features. *Advances in Intelligent Systems and Computing* 868, 150–159 (2019), cited By 0
5. Hussain, Z., Gimenez, F., Yi, D., Rubin, D.: Differential data augmentation techniques for medical imaging classification tasks. In: *AMIA Annual Symposium Proceedings*. vol. 2017, p. 979. American Medical Informatics Association (2017)
6. Jafari, M.H., Karimi, N., Nasr-Esfahani, E., Samavi, S., Soroushmehr, S.M.R., Ward, K., Najarian, K.: Skin lesion segmentation in clinical images using deep learning. In: *2016 23rd International Conference on Pattern Recognition (ICPR)*. pp. 337–342 (Dec 2016)
7. Jafari, M.H., Nasr-Esfahani, E., Karimi, N., Soroushmehr, S.M.R., Samavi, S., Najarian, K.: Extraction of skin lesions from non-dermoscopic images for surgical excision of melanoma. *International Journal of Computer Assisted Radiology and Surgery* 12(6), 1021–1030 (Jun 2017), <https://doi.org/10.1007/s11548-017-1567-8>
8. Jerant, A.F., Johnson, J.T., Sheridan, C., Caffrey, T.J.: Early detection and treatment of skin cancer. *American family physician* 62(2) (2000)
9. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1*. pp. 1097–1105. NIPS’12, Curran Associates Inc., USA (2012), <http://dl.acm.org/citation.cfm?id=2999134.2999257>
10. Litjens, G., Kooi, T., Bejnordi, B.E., Setio, A.A.A., Ciompi, F., Ghafoorian, M., van der Laak, J.A., van Ginneken, B., Snchez, C.I.: A survey on deep learning in medical image analysis. *Medical Image Analysis* 42, 60 – 88 (2017), <http://www.sciencedirect.com/science/article/pii/S1361841517301135>
11. Nachbar, F., Stolz, W., Merkle, T., Cognetta, A.B., Vogt, T., Landthaler, M., Bilek, P., B.-Falco, O., Plewig, G.: The abcd rule of dermatoscopy: high prospective value in the diagnosis of doubtful melanocytic skin lesions. *Journal of the American Academy of Dermatology* 30(4) (1994)
12. Nida, N., Irtaza, A., Javed, A., Yousaf, M., Mahmood, M.: Melanoma lesion detection and segmentation using deep region based convolutional neural network and fuzzy c-means clustering. *International Journal of Medical Informatics* 124, 37–48 (2019)
13. Ruela, M., Barata, C., Marques, J., Rozeira, J.: A system for the detection of melanomas in dermoscopy images using shape and symmetry features. *Computer Methods in Biomechanics and Biomedical Engineering: Imaging and Visualization* 5(2), 127–137 (2017)
14. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. vol. 00, pp. 1–9 (June 2015), [doi.ieeecomputersociety.org/10.1109/CVPR.2015.7298594](https://doi.org/10.1109/CVPR.2015.7298594)
15. Victor, A., Ghalib, M.: Automatic detection and classification of skin cancer. *International Journal of Intelligent Engineering and Systems* 10(3), 444–451 (2017)
16. Yadav, V., Kaushik, V.: Detection of melanoma skin disease by extracting high level features for skin lesions. *International Journal of Advanced Intelligence Paradigms* 11(3-4), 397–408 (2018)
17. Yu, L., Chen, H., Dou, Q., Qin, J., Heng, P.A.: Automated melanoma recognition in dermoscopy images via very deep residual networks. *IEEE Transactions on Medical Imaging* 36(4), 994–1004 (2017)